Applying the Standardized Process for Data Analytics

Flocon 2024

Mobile, Alabama

January 9, 2024

Matthew Spitzer, PhD; Rosalie Bakken, PhD; and Jennifer Marr, BA

Scenario 1

Threat: Third-party breach using vended equipment on the network; C&C; risks from excessively large attack surface

Hypothesis: Devices on the network are creating connections that are unexpected and not in line with the vendor’s documentation of standard device behavior.

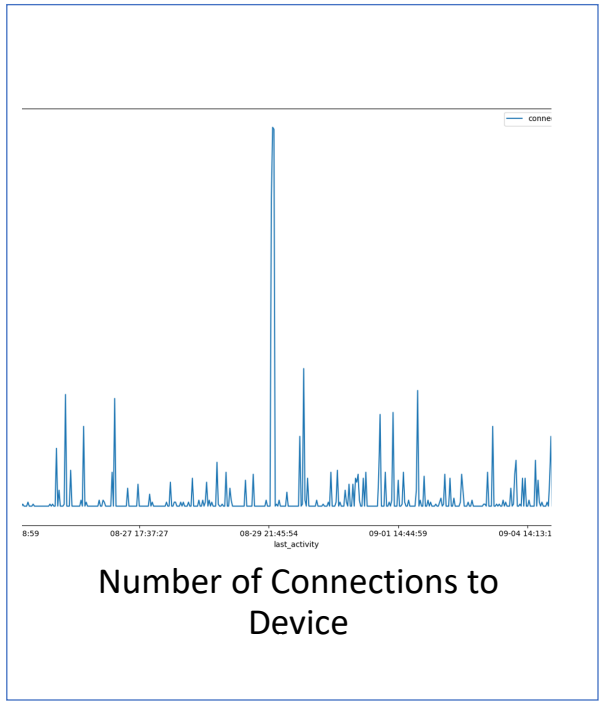
Scenario 1: Background and Logic

- Background
  - A select group of new devices were observed making unexpected connections after connecting to the network
  - Problem to be addressed:
    - Confirm the observed activity that initially created an alert
    - Pinpoint any further activity that was unexpected
- Steps
  - Create the map
    - Obtain as much detail as possible on the devices and their activity
  - Describe our destination
    - Retrieve all connection activity associated with the devices
  - Map a course to the destination
    - Analyze the connection activity for patterns and outliers

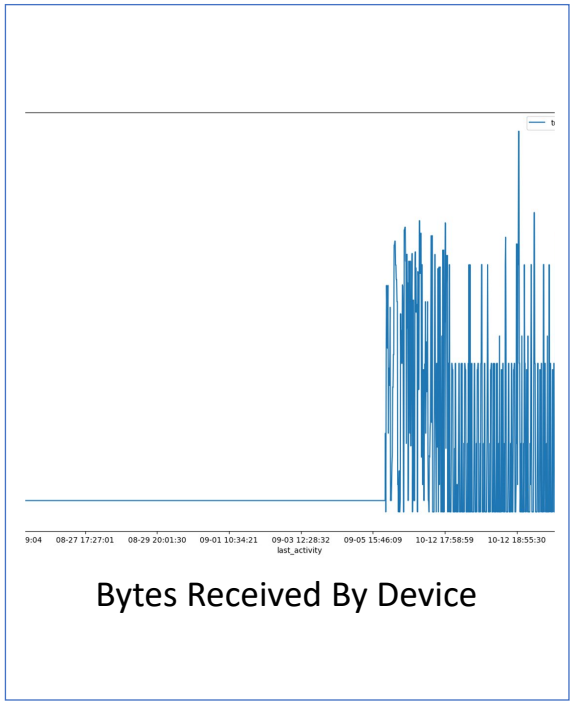
Scenario 1: Data and Analysis

- High-Level Logic Summary
  - Obtain start and end dates of the devices' time on the network
  - Include time-sensitive resolution of historically accurate IP addresses due to dynamic assignments
  - Select all connections where the IP addresses were associated with the devices, regardless if the IP addresses were the source or destination of the connection
  - Data sources: Netflow, DHCP, DNS
- Launch the Boat
  - Volumes of connections
  - Breakdown of connections by whether the device was the source or destination of the connection
    - Internal and external IP addresses involved
    - Unique ports and frequencies
    - Byte volumes
    - Connection durations
    - Key question: is the connection expected, according to the vendor documentation?
  - Remote access connections
    - Patterns by timeframe
    - Patterns by volumes
    - Patterns by external IP addresses involved

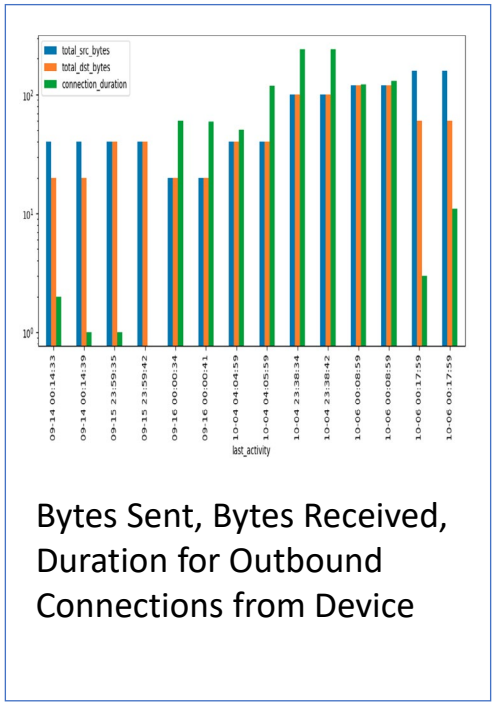
Scenario 1: Statistical Analysis Results



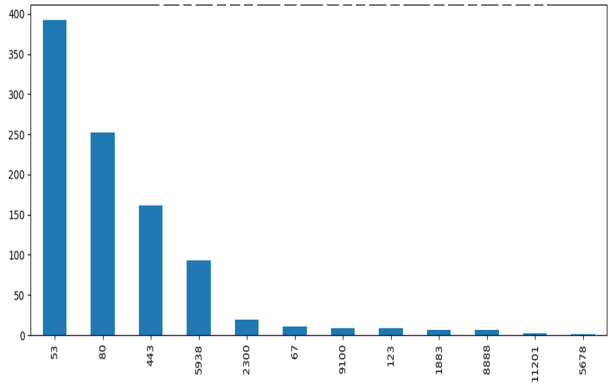
Number of Connections to Device



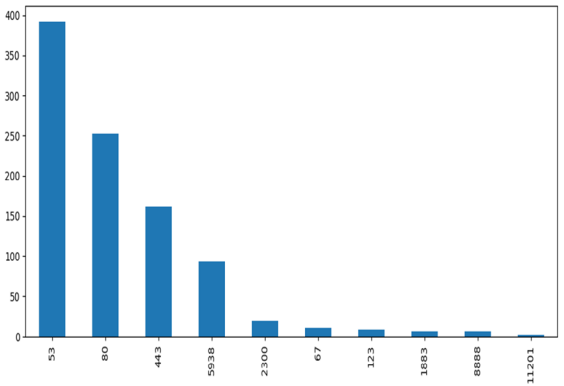
Bytes Received By Device



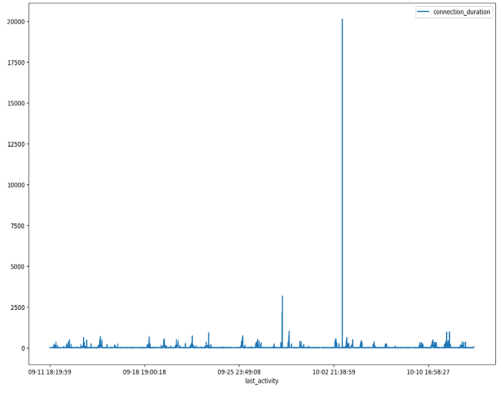
Bytes Sent, Bytes Received, Duration for Outbound Connections from Device



Port Frequencies for Inbound Connections to Device



Port Frequencies for Outbound Connections from Device



Durations for Outbound Connections from Device

Scenario 1: Detours and Nuances

- Detours and shiny objects
  - What was the content of the data was transferred in the remote connection activity?
  - Why were the devices connecting to unexpected cloud services?
  - Odd patterns in bytes sent/received by the devices
  - Guest wireless network usage
  - Bridged connectivity through both wireless and cellular networks
- Nuances
  - How does one define a “successful” connection for this hypothesis?
  - Data source availability may prohibit result creation in certain scenarios
  - Do policies exist to aid in enforcement for unapproved connections?

Scenario 2

Threat: Enterprise users are utilizing easy-to-guess passwords; risks associated with credential hijacking; Living off the Land techniques

Hypothesis: A password spray exercise performed by the Red Team can be detected in login failure data. Additionally, acquiesced user accounts can be identified.

Scenario 2: Background and Logic

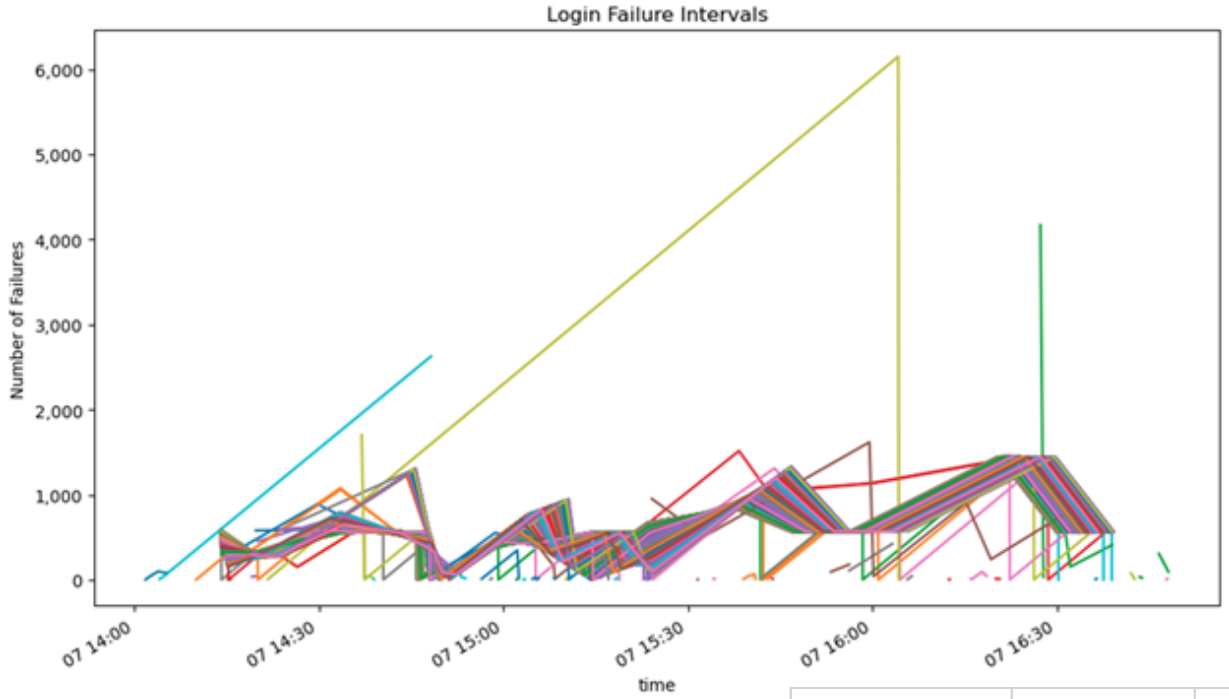
- Background
  - A password spray test was conducted by the Red Team in support of a proactive threat response process creation
  - Problem to be addressed
    - Identify the password spray test occurred
    - Identify any user accounts compromised during this internal exercise
- Steps
  - Create the map
    - Understand the parameters of the password spray test
    - Gain familiarity with login failure data
  - Describe the destination
    - Recognize failure patterns related to repeated authentication attempts on the same accounts
      - Not all authentication failures are related to password spray attempts
      - Realize jitter can affect password spray interval detection
  - Map a course to the destination
    - Retrieve all login failure data points for all accounts
    - Statistically analyze patterns between repeated failures on multiple occurrences of the same credentials

Scenario 2: Data and Analysis

- High-Level Logic Summary
  - Retrieve login failure data from the web-based authentication mechanisms
  - Acquire definitive start and end dates of the Red Team testing effort
  - Data sources: Active Directory Federated Services
- Launch the Boat
  - Analysis efforts
    - Group failed logins by user
    - Calculate intervals between failures for each account
    - Map intervals
    - ML Modeling procedures

Scenario 2: Statistical Analysis Results and Modeling

Obtain results and correct course as needed



Course Correction

Several machine learning models were attempted to find compromised accounts based solely on the failure data:

- First model – DBSCAN. The data did not provide the continuous values needed to support a DBSCAN clustering effort; the data points were sufficiently dissimilar to prevent effective clustering.
- Second model – KMeans clustering. Results were dissimilar and did not include continuous variables, preventing effective clustering.
- Third model – LCA. The third model was attempted based on identifying other clustering techniques that work more effectively on categorical rather than continuous data.

Latent_class	user	label
0	abc123	36
1	**abc102**	1
1	**abc103**	1
2	**abc102**	33
2	**abc103**	33
2	abc106	34
2	abc107	34
2	abc108	36
2	abc109	34
2	abc110	32
2	abc112	34
2	abc113	34
2	abc114	17

Scenario 2: Nuances

Arrival at Destination

- Nuances
  - What constitutes a statistically significant interval pattern of login failures?
  - Lack of certain data (success logs) may prevent identifying the broader scope and downstream activity
  - Understanding the nature of the data (categorical vs. continuous) can aid in machine learning model selection
  - LCA model provided the best fit for categorical data
    - Some results may prove challenging during the interpretation phase

Latent_class	user	label
0	abc123	20
1	abc123	10
2	abc123	6

Questions?